



Petabyte-Scale Computing for High Energy Physics

Stephen Wolbers, Fermilab

ACM Chicago
October 17, 2001



Outline

- Introduction to High Energy Physics (HEP)
- Computing Issues in High Energy Physics
 - How are computing and physics related?
 - Data, data, data
 - Access to and analysis of the data
 - Software Engineering in a collaborative environment
- New Ideas in Computing and HEP
 - GRID Computing
 - “Tape Killer”
- Summary



Introduction to HEP

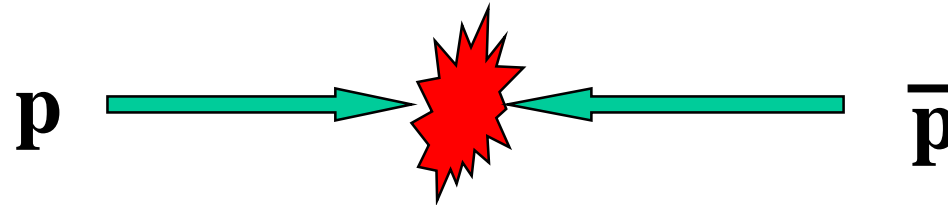
- Particle physics uses beams of particles striking targets to study the fundamental nature of matter and interactions.
- Advances in the field have come from:
 - Higher Energy Particles and Interactions (big accelerators, cosmic rays)
 - More collisions per unit time and space (luminosity)
 - Better detectors
 - More sensitivity, more granular, fewer cracks, lower deadtime, more radiation-hard
 - More “events” saved to storage (disk or tape)
 - More sophisticated analysis of “events”
 - Better simulation of the beams, collisions, and detector
 - Advances in Theory

Computing

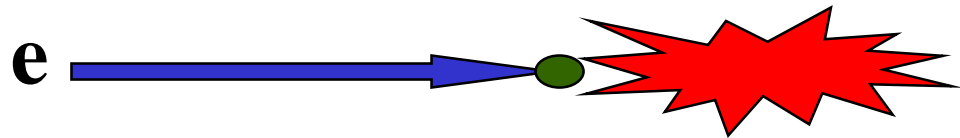


Collisions Simplified

- Collider:

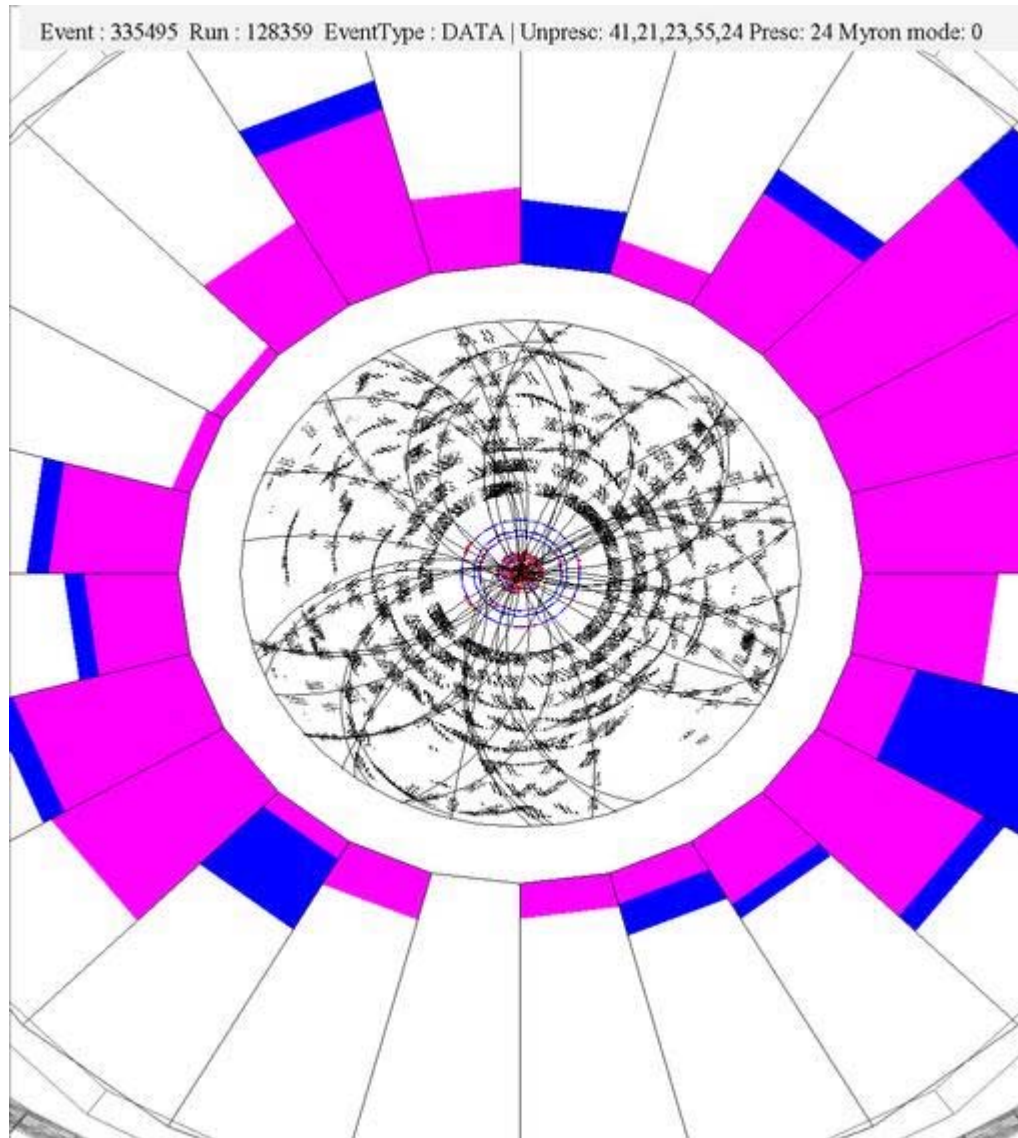


- Fixed-Target:





CDF Collisions

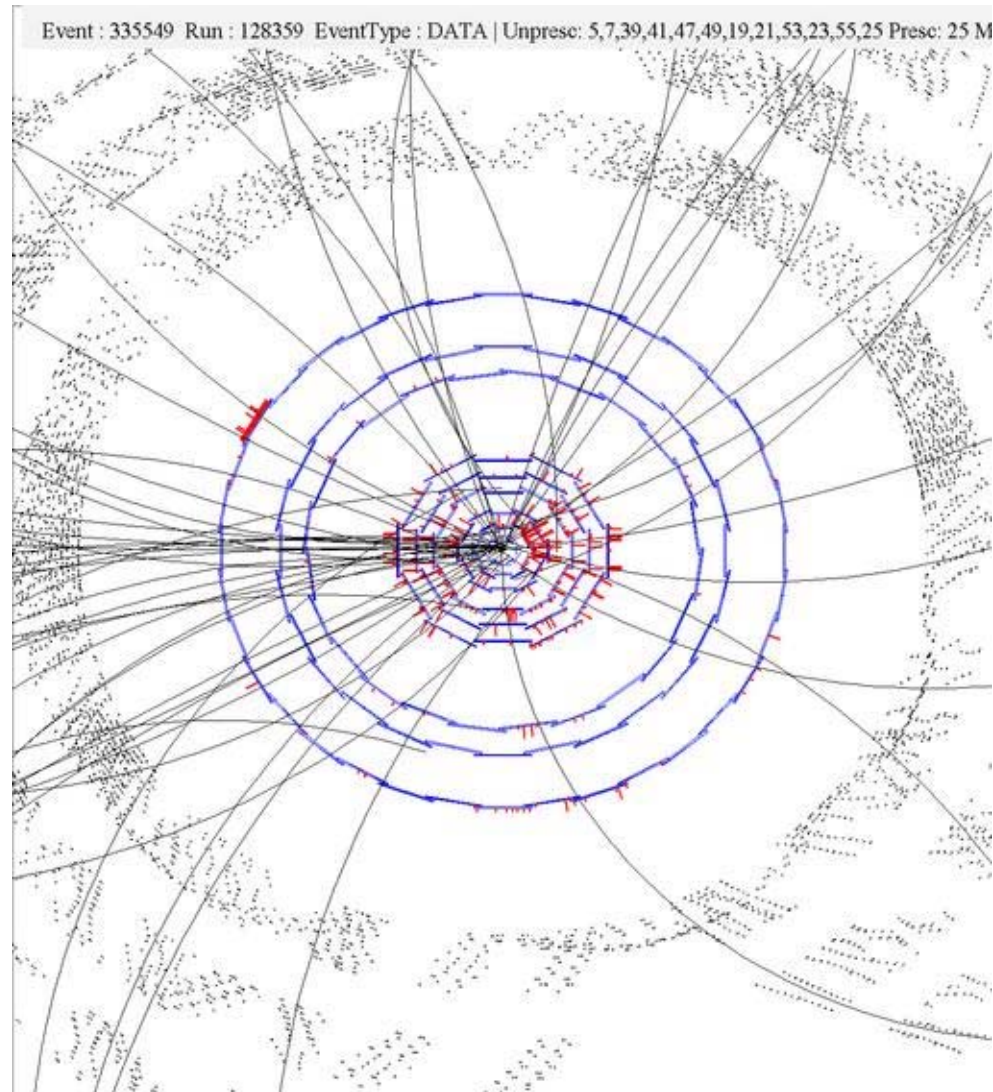


October 17, 2001

Stephen Wolbers ACM Chicago
October 2001



CDF Collisions

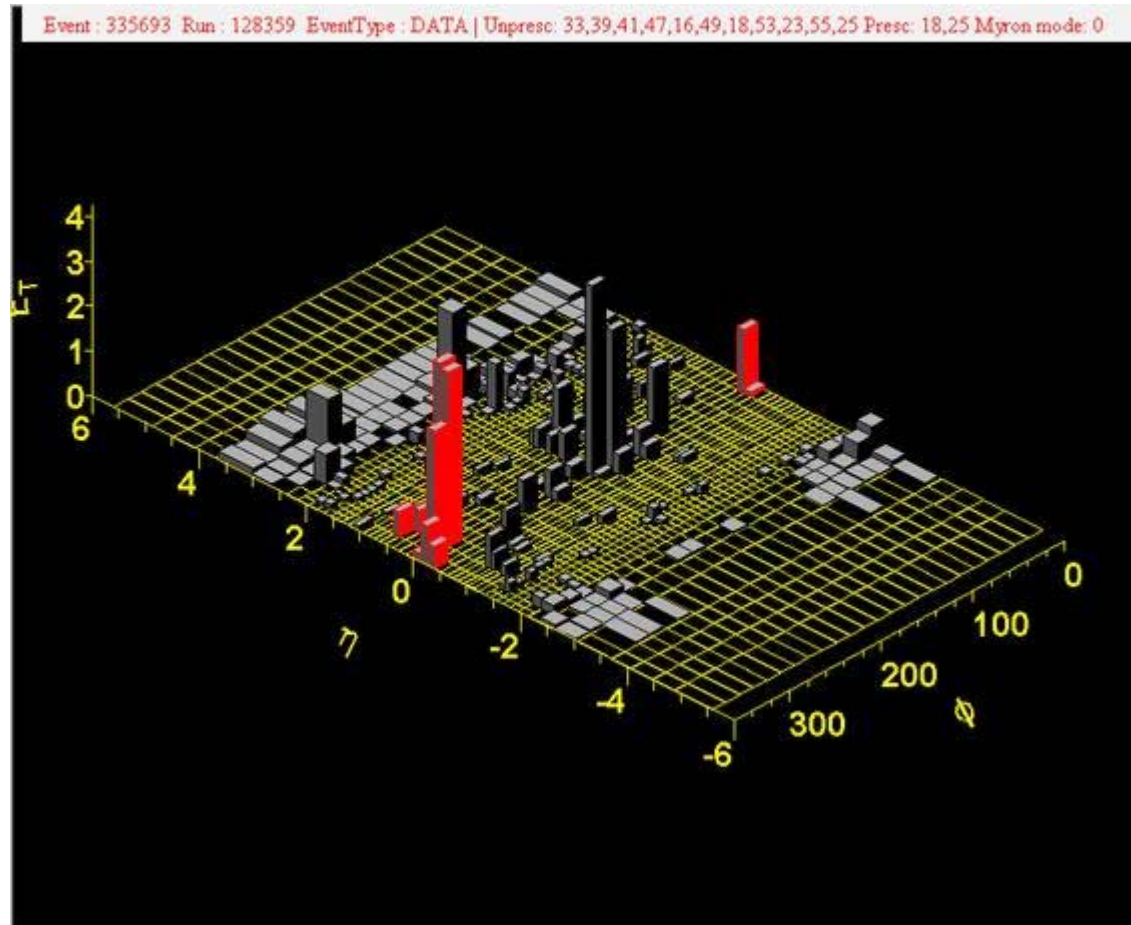


October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

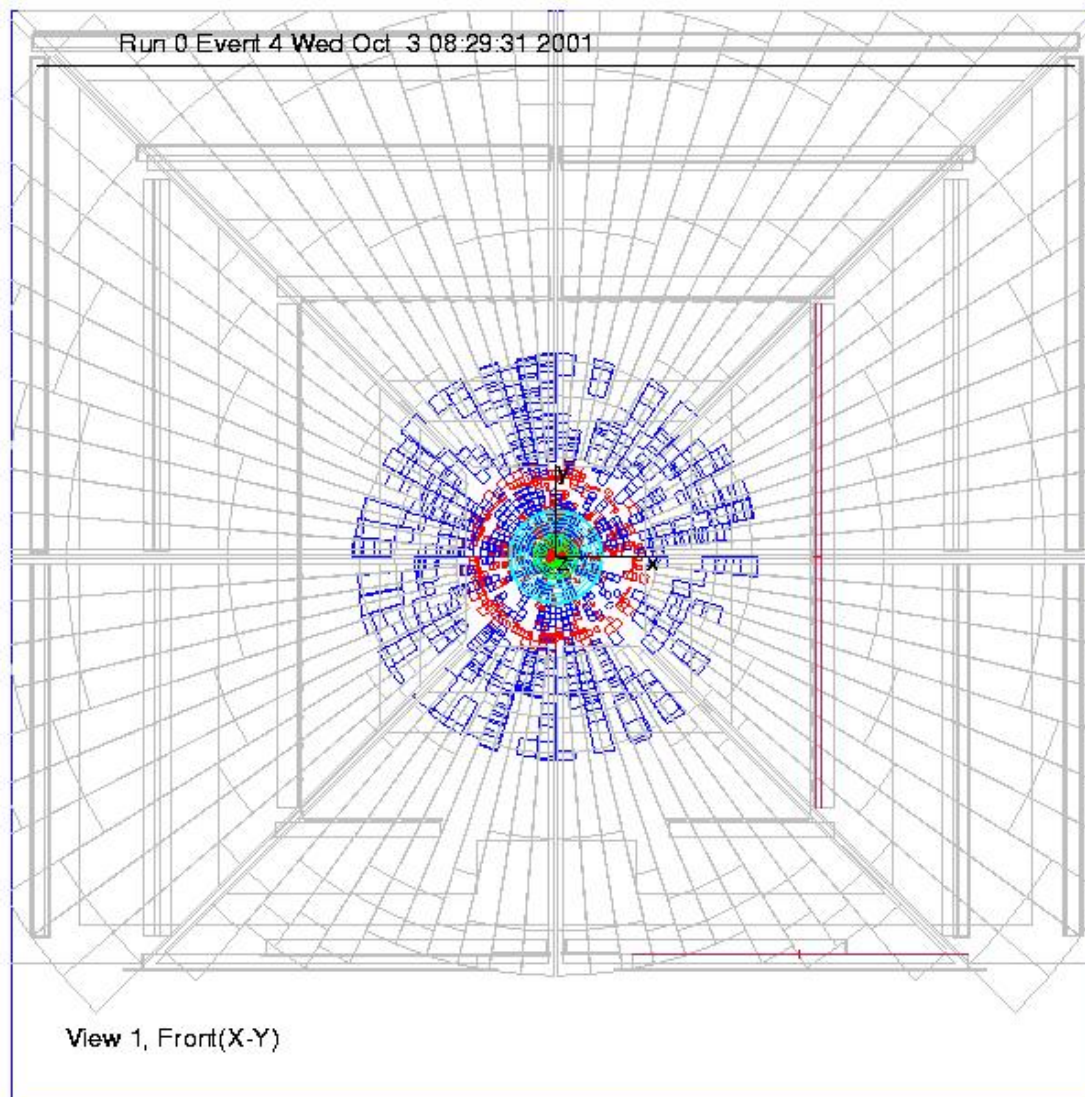


CDF Collisions





D0 Collisions



October 17, 2001

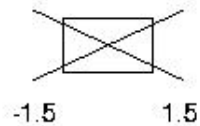
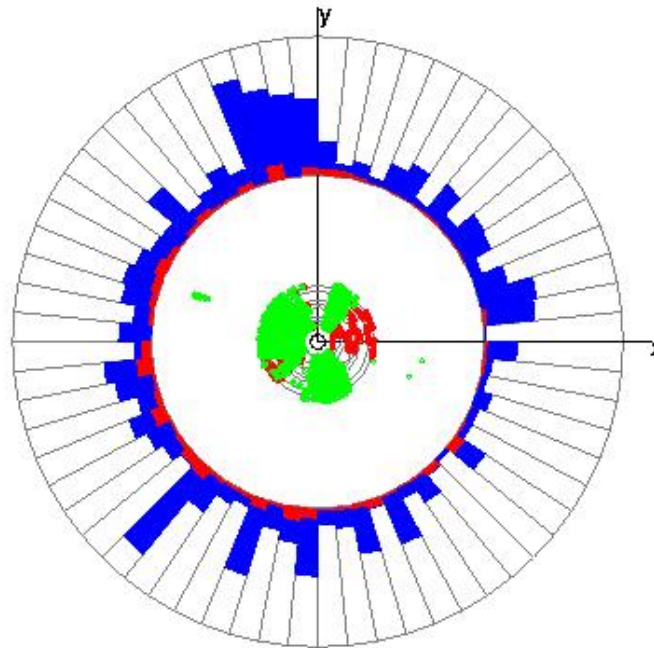
Stephen Wolbers ACM Chicago
October 2001



D0 Collisions

Run 0 Event 4 Wed Oct 3 08:45:21 2001

ET scale: 3 GeV



October 17, 2001

Stephen Wolbers ACM Chicago
October 2001



Particle Acceleration

- Particle acceleration occurs in a multi-step process
- At Fermilab (Fermi National Accelerator Laboratory), Batavia, Illinois
 - Cockcroft-Walton accelerator
 - Linac (Linear accelerator)
 - Booster
 - Main Injector
 - Tevatron
 - Anti-Protons
 - Accumulator/debuncher
 - Recycler



October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

11



Cockcroft-Walton



October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

12



Linac



October 17, 2001

Stephen Wolbers ACM Chicago
October 2001



Booster



Main Injector



Tevatron



Antiproton Source



Recycler



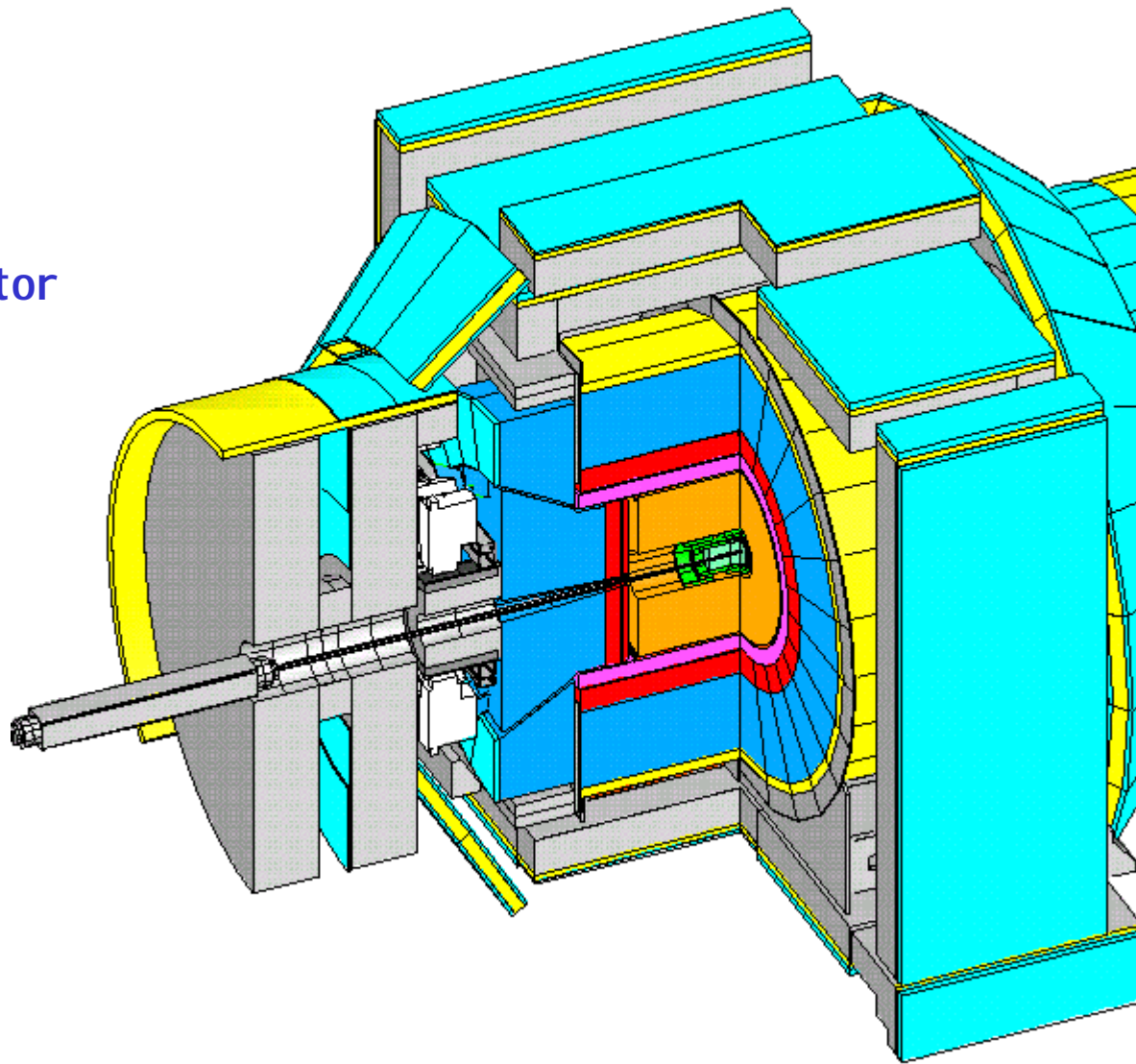


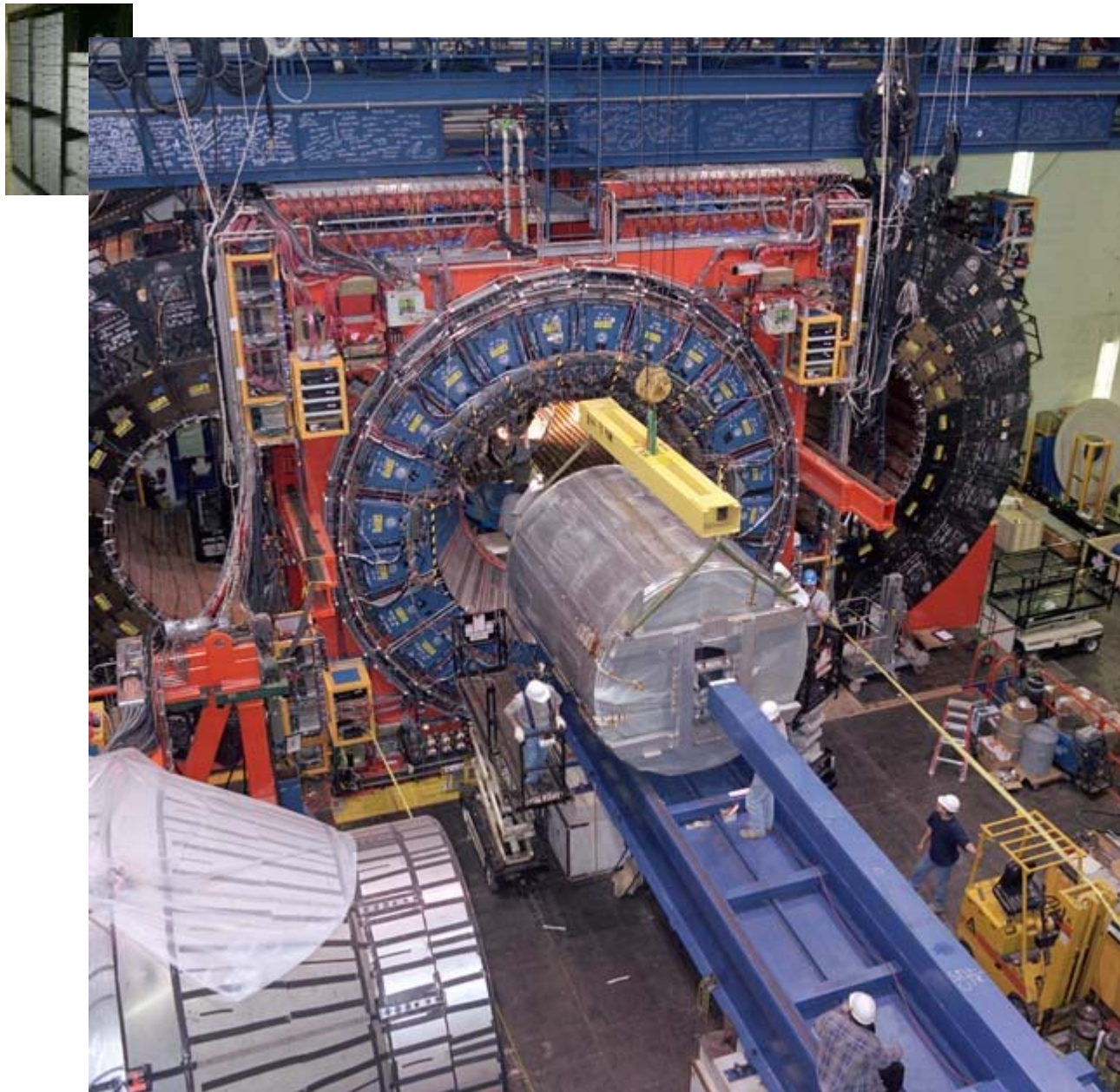
Particle Detectors

- Large detectors are used to “see” the interactions of particles.
- Consist of subdetectors which record information about particle position, energy and momentum.
 - Used to measure the number, types and properties of particles coming from collisions.
 - Also used to identify particle decays.
 - e , μ , π , γ , p , K , W , Z , b , c , ...
- Built by many institutions and by many collaborators.
 - CDF, large detector at Fermilab, has over 500 physicists from over 50 institutions from all over the world.



CDF Detector





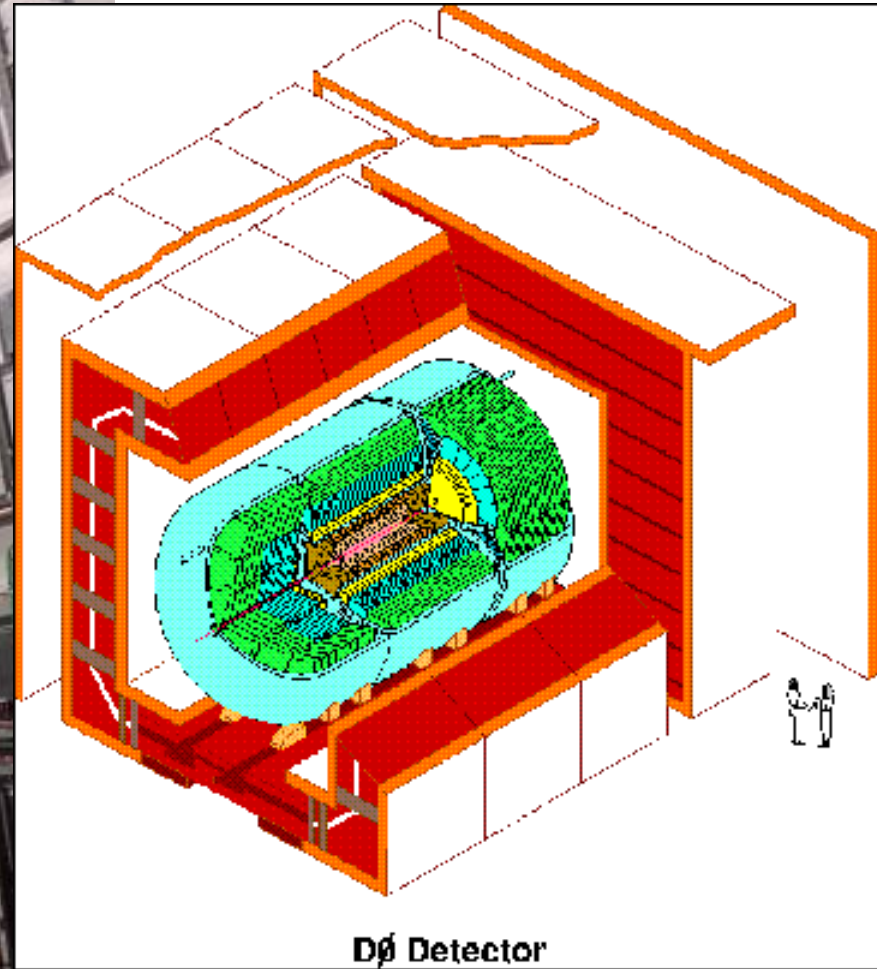
CDF Particle Detector

October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

18

D0 Particle Detector



October 17, 2001

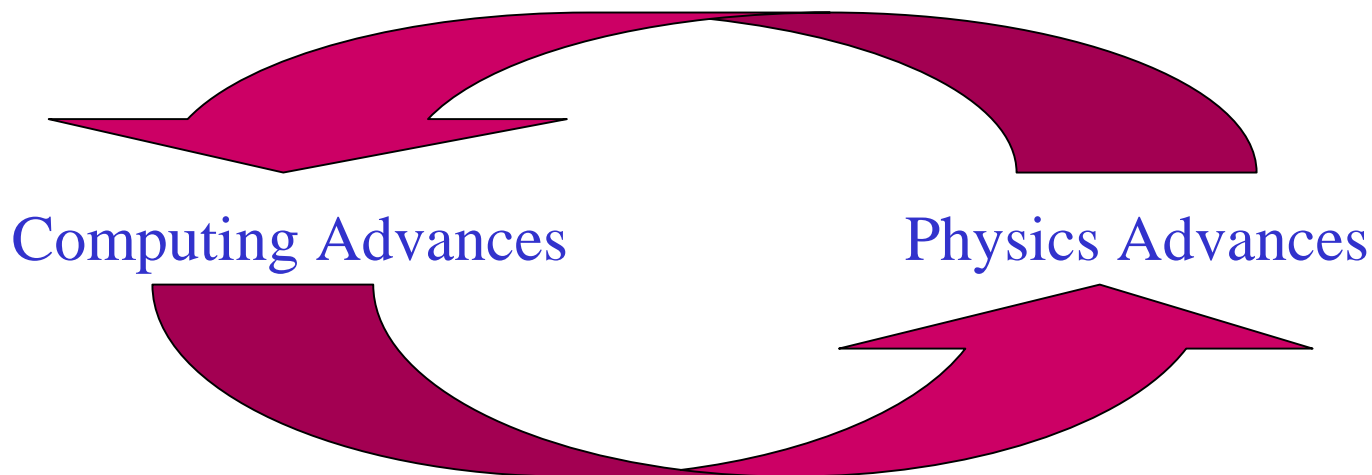
Stephen Wolbers ACM Chicago
October 2001

19



Computing and Particle Physics

- HEP has always required substantial computing resources
 - Computing advances have enabled “better physics”.
 - Physics research demands further computing advances.
 - Physics and computing have worked together over the years.





In Silica Fertilization

All Science Is Computer Science

By GEORGE JOHNSON

EXCEPT for the fact that everything, including DNA and proteins, is made from quarks, particle physics and biology don't seem to have a lot in common. One science uses mammoth particle accelerators to explore the subatomic world; the other uses petri dishes, centrifuges and other laboratory paraphernalia to study the chemistry of life. But there is one tool both have come to find indispensable: supercomputers powerful enough to sift through piles of data that would crush the unaided mind.

Last month both physicists and biologists made announcements that challenged the tenets of their fields. Though different in every other way, both discoveries relied on the kind of intense computer power that would have been impossible to marshal just a few years ago. In fact, as research on so many fronts is becoming increasingly dependent on computation, all science, it seems, is becoming computer science.

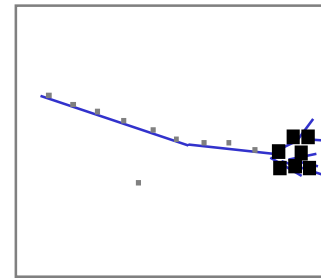
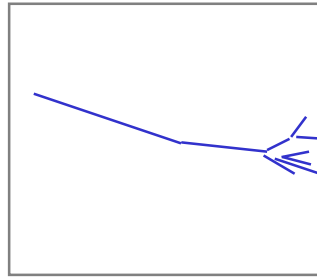
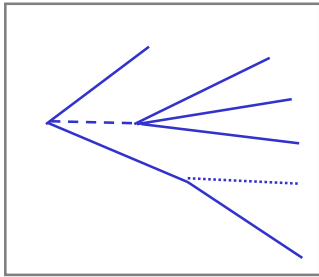
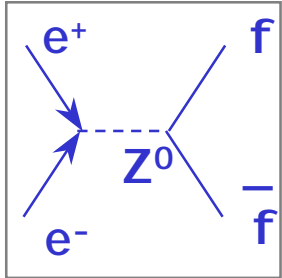
"Physics is almost entirely computational now," said Thomas B. Kepler, vice president for academic affairs at the Santa Fe Institute, a multidisciplinary research center in New Mexico. "Nobody would dream of doing these big accelerator experiments without a tremendous amount of computer power to analyze the data."

New York Times,
Sunday, March 25, 2001

October 17, 2001



Physics to Raw Data (taken from Hans Hoffman, CERN)



```
2037 2446 1733 1699
4003 3611 952 1328
2132 1870 2093 3271
4732 1102 2491 3216
2421 1211 2319 2133
3451 1942 1121 3429
3742 1288 2343 7142
```

"Nature"

Fragmentation,
Decay

Interaction with
detector material
Multiple scattering,
interactions

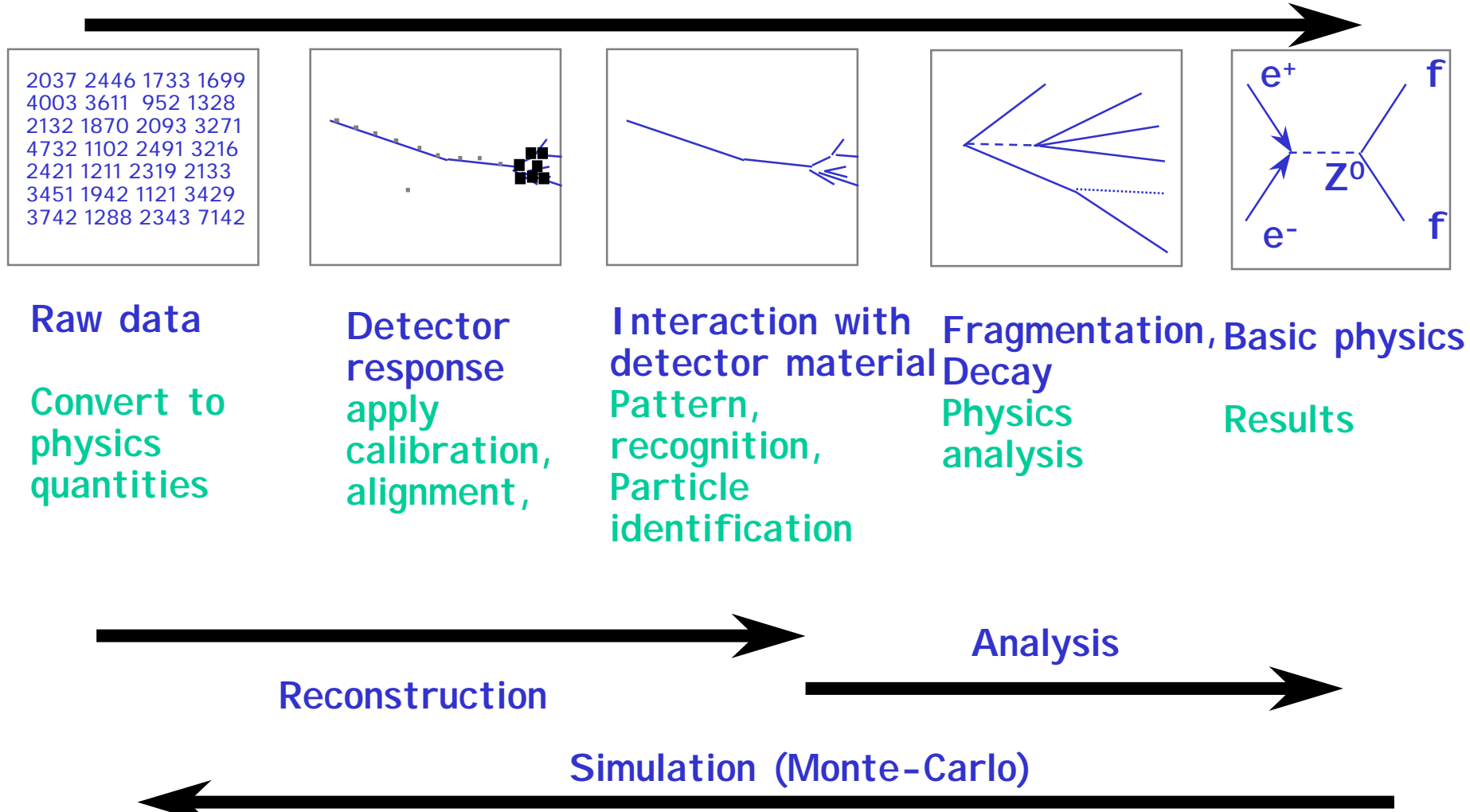
Detector
response
Noise, pile-up,
cross-talk,
inefficiency,
ambiguity,
resolution,
response
function,
alignment,
temperature

Raw data
(Bytes)

Read-out
addresses,
ADC, TDC
values,
Bit patterns



From Raw Data to Physics





Computing Connection

Desired Improvement

Computing Technique

Higher energy	→	Accelerator Design/simulation
More collisions	→	Acc. Design and controls
Better detectors	→	Triggers (networks, CPU), simulation
More events	→	Disk, tape, CPU, networks
Better analysis	→	Disk, tape, CPU, networks, algorithms
Simulation	→	CPU, algorithms, OO
Theory	→	CPU, algorithms, OO



Data, Data, Data

- HEP has always been very data-intensive.
- To advance the science, more data is required:
 - To search for rarer phenomena
 - To study phenomena in more detail
 - To study more and different types of phenomena
- The science is truly “data-limited”.
- More data invariably leads to more science.



Data Volumes for HENP Experiments

(in units of 10^9 bytes)

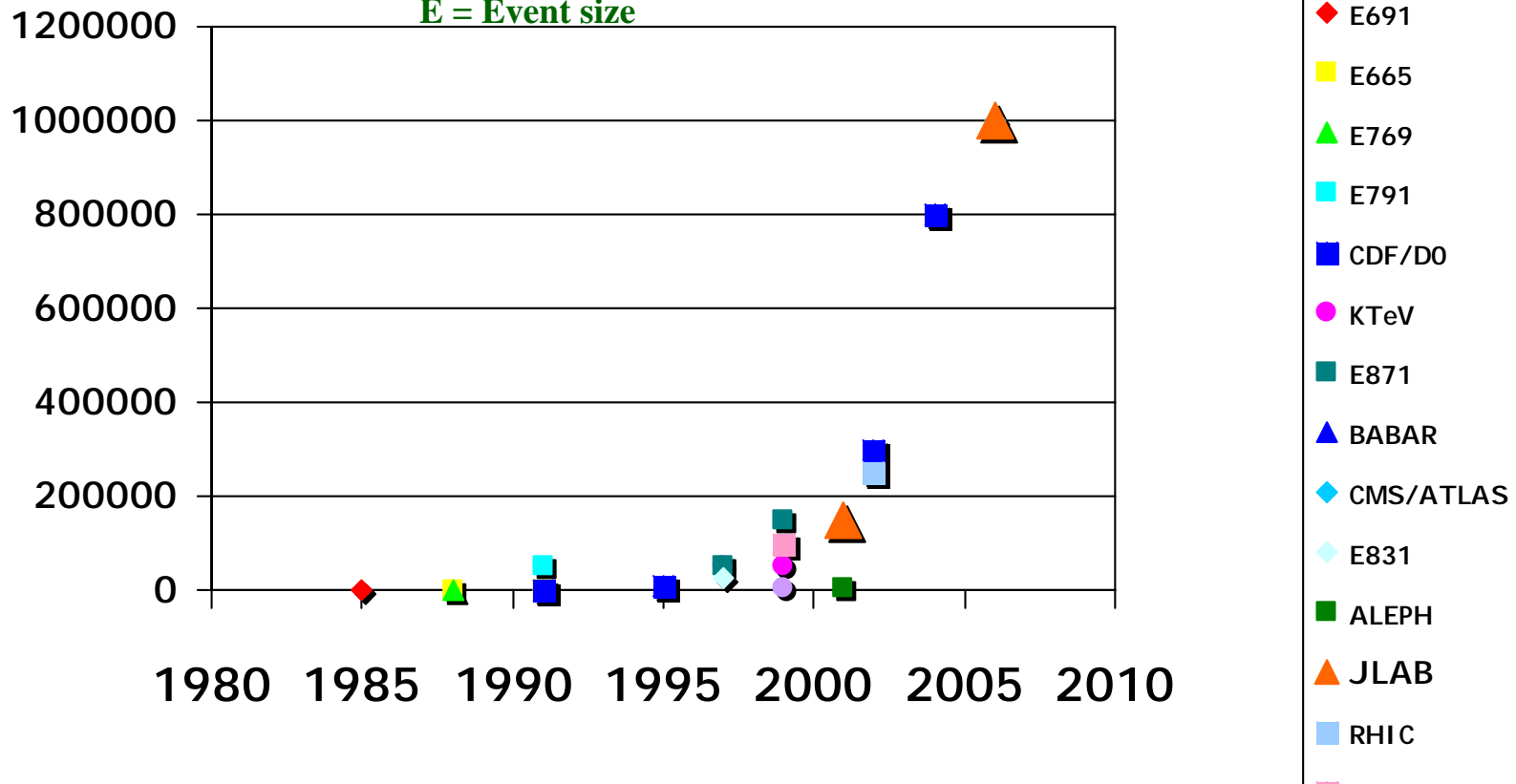
$$\text{Volume} = \sigma L \varepsilon E$$

σ = Cross Section

L = Luminosity

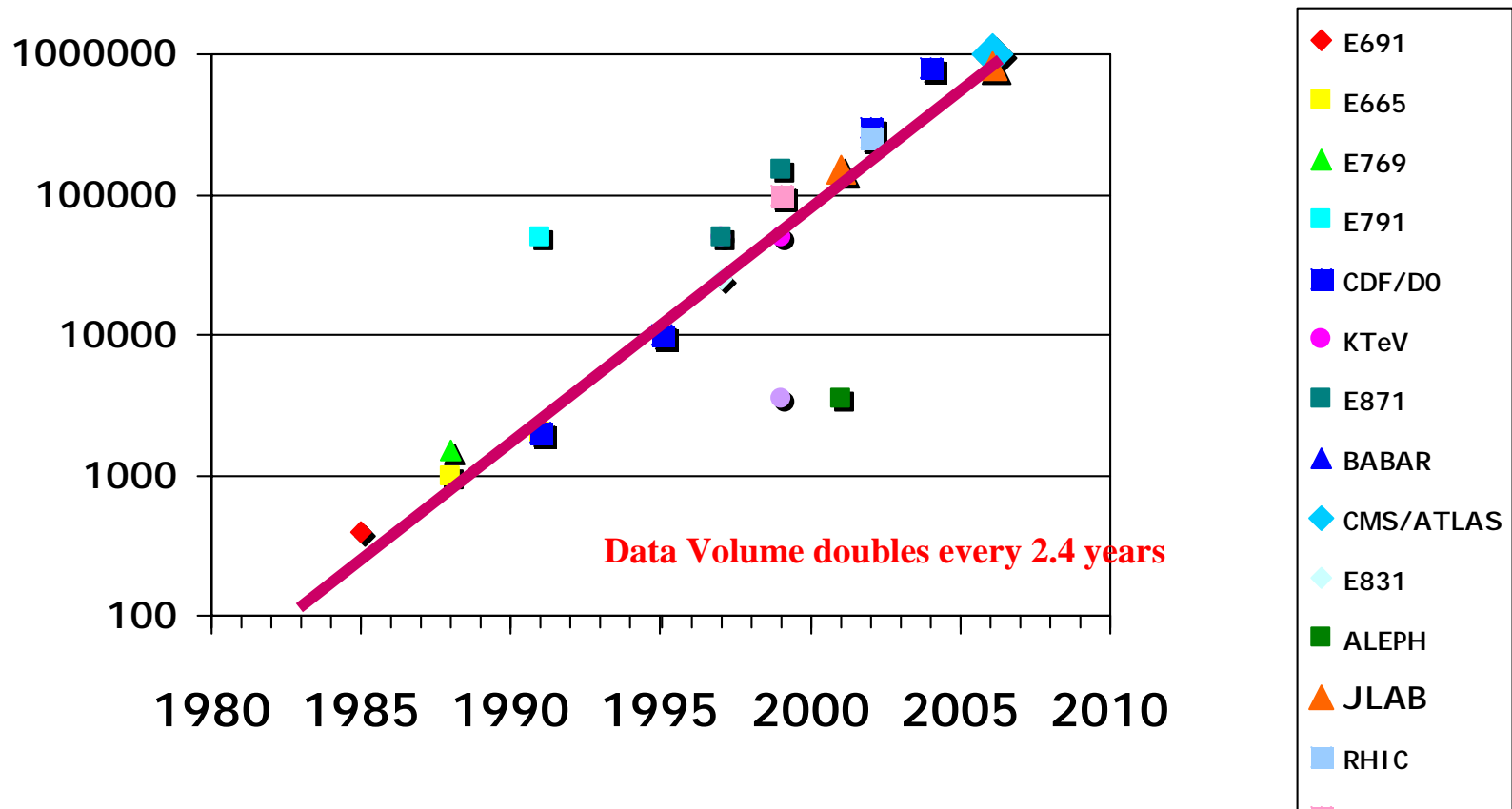
ε = Efficiency for collecting data

E = Event size





Data Volume per experiment per year (in units of 10^9 bytes)





Computing at Fermilab for the Collider Run 2 (Run 2a began March 1, 2001)

- **Data Volume and Rate!**
 - Data volume of ~1 Pbyte (over 2 years, CDF+D0)
(Pbyte = 10^{15} bytes = 1,000,000 Gbytes)
 - Typical hard disk is 20 Gbytes, DVD is 5 Gbytes.
 - 1 Pbyte is 200,000 DVD-equivalents.
 - Rates out of detector to storage up to 20 Mbyte/sec (each detector).
 - Equivalent to very good digital video.
- **Substantial CPU power is required to analyze all of this data.**
 - Recording the data is only the first step.
 - Making the data available to all users and computing tasks is essential!



The Future Holds Much More Data!

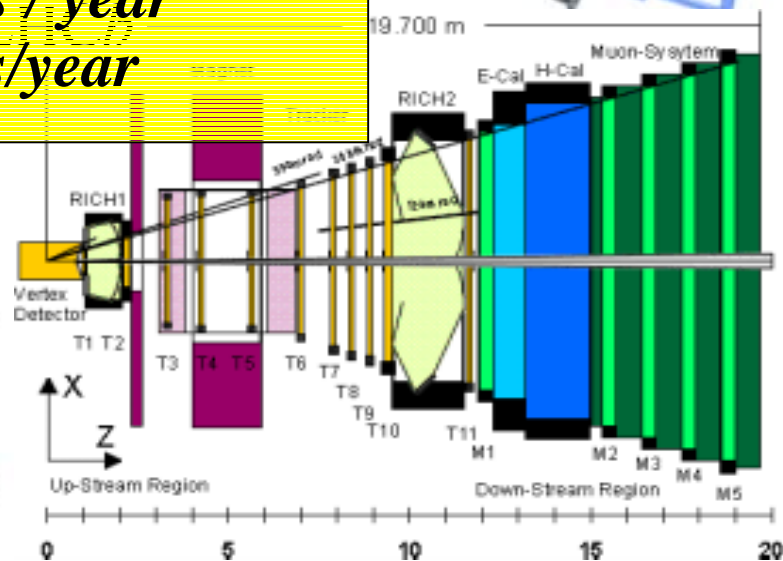
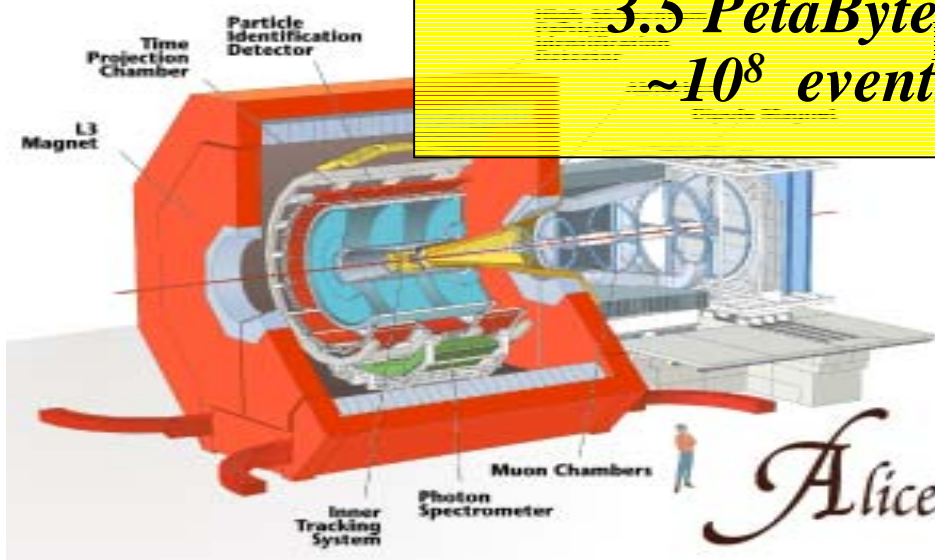
- Run 2b at Fermilab will run in 2003-2007.
 - Data rates will be larger, maybe much larger.
- The next big accelerator is being constructed at CERN in Geneva, Switzerland.
 - It will collide protons at 7x the energy of the Fermilab accelerator.
 - The total data-taking rates and volumes are truly spectacular.
 - Total tape storage : 28.5 Petabytes/year
 - Total disk storage : 10.4 Petabytes/year
 - Total CPU : 7.349 Million SpecInt95
 - Network : 4.810 Gbps WAN bandwidth

The LHC Detectors

ATLAS

CMS

*Raw recording rate 0.1 – 1 GB/sec
3.5 PetaBytes / year
~10⁸ events/year*



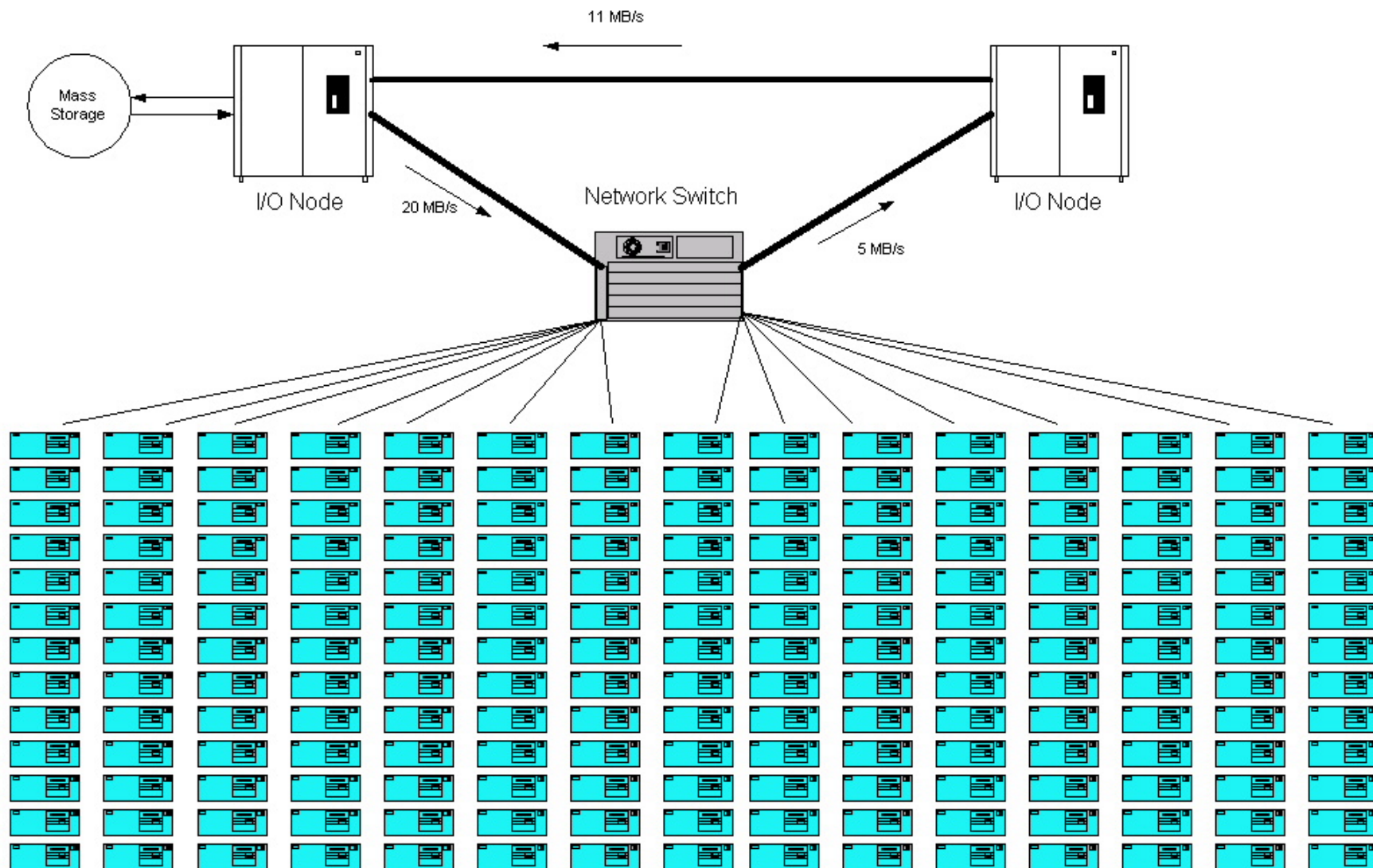


Run 2 Computing at Fermilab: How are the data processed and accessed?

- Computing power to process the data:
 - Large arrays (100's) of PCs ("farms")
 - Each event can be analyzed independently of all others, making the problem trivially parallel.
 - To handle the full rate of 75 events/sec with each event taking 5 seconds on a PIII/500 processor, one needs:
 - 375 500-MHz processors, or the equivalent.



Run II CDF PC Farm





October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

33



October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

34



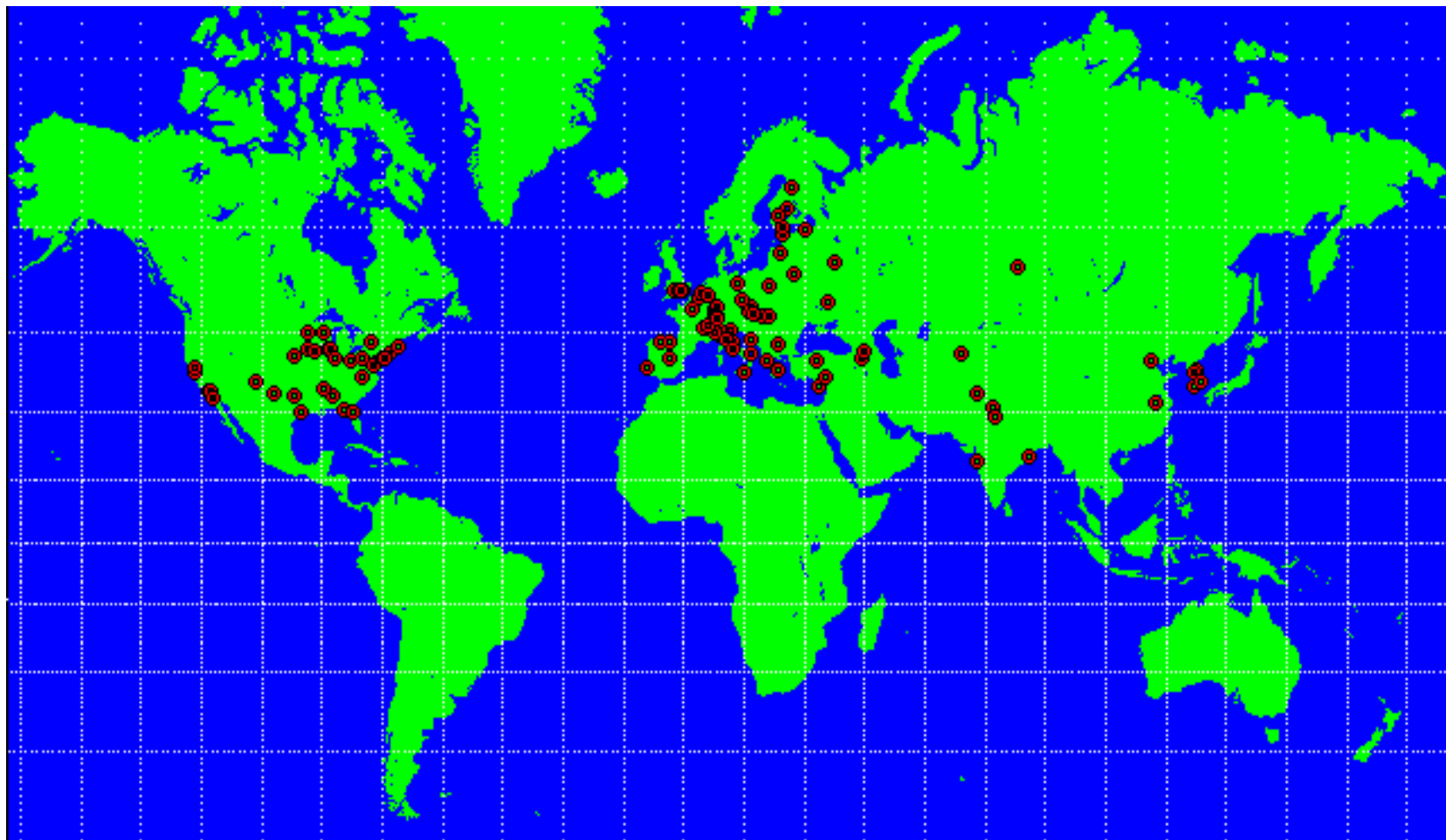
Accessing Data and Analyzing I t

- Processed Data is Stored on Disk (whenever possible) and in Tape Robots.
- Access is directly from disk (large SMP) or over the network.
- This is an area of active investigation
 - Arrays of PCs
 - Fibrechannel
 - Network accessible data
 - Etc.
- Complicating this whole system is the fact that collaborators live all over the world and want to analyze the data wherever they may be.



World Wide Collaboration

⇒ distributed computing & storage capacity





BaBar: Worldwide Collaboration of 80 Institutes



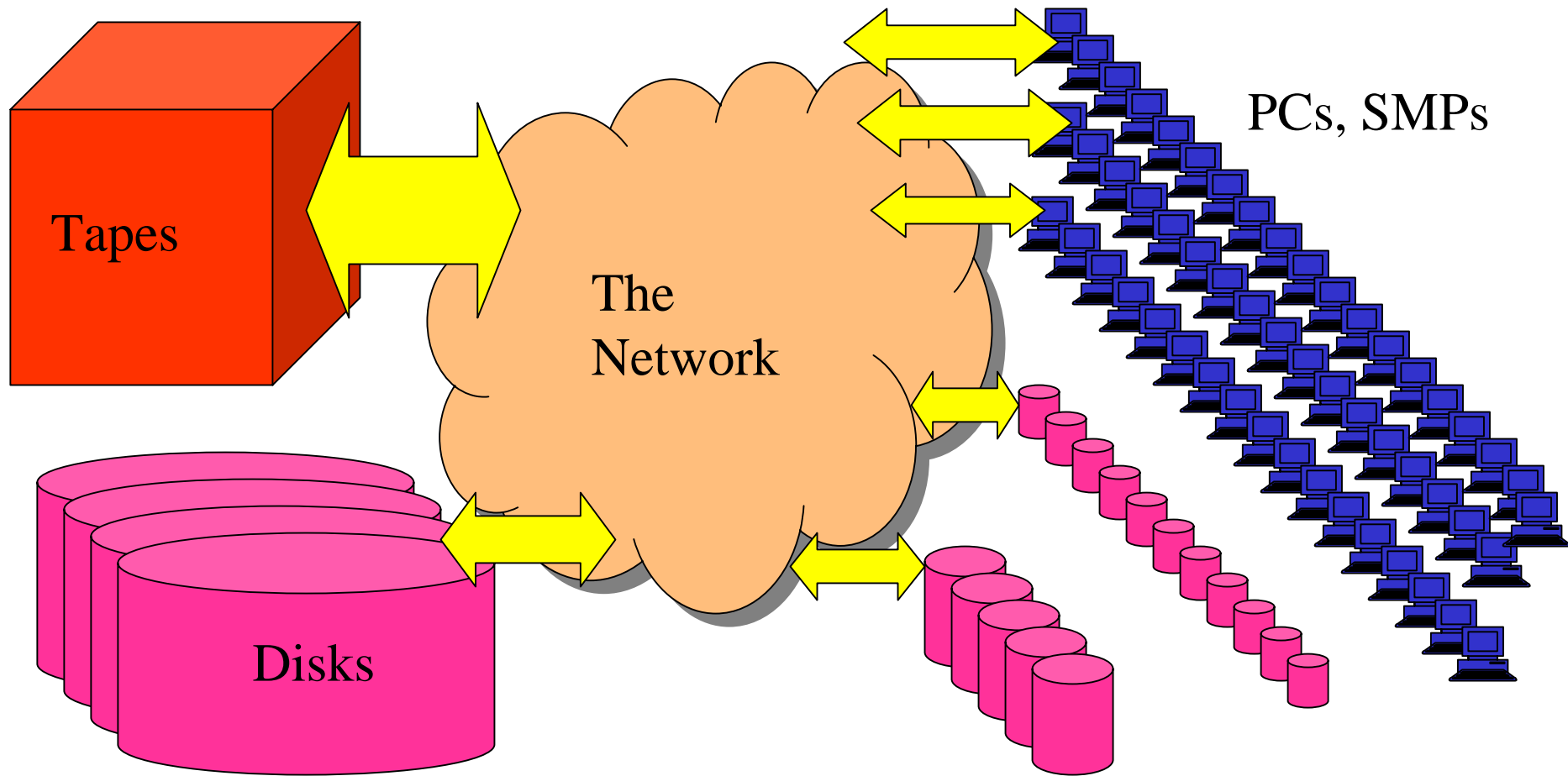
October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

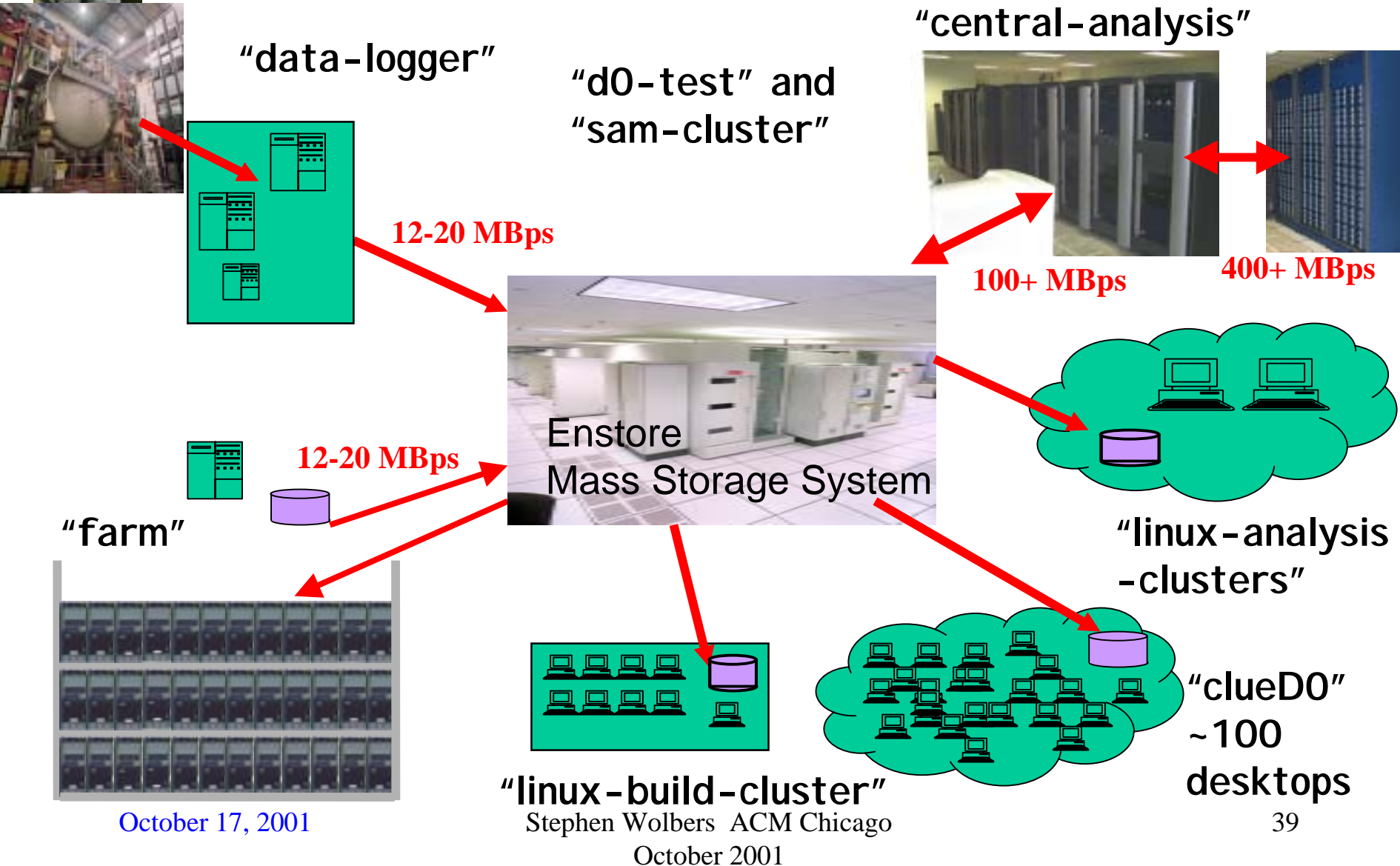
37



Analysis – a very general model



D0 computing systems

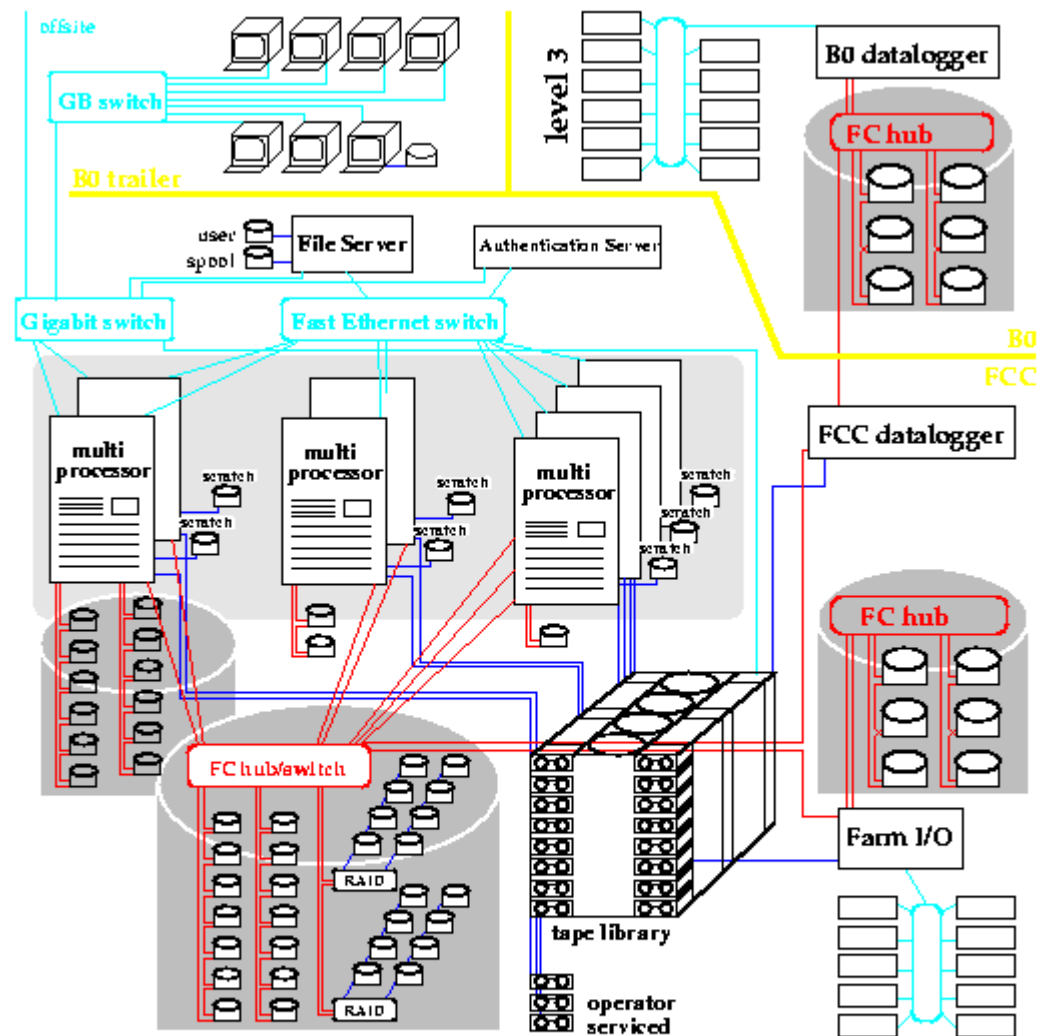




Data Access Model: CDF

Ingredients:

- Gigabit Ethernet
- Raw data are stored in tape robot located in FCC
- Multi-CPU analysis machine
- High tape access bandwidth
- Fiber Channel connected disks





Software Development in a collaborative environment

- **Event Reconstruction Software**
 - Written by physicists.
 - Translates detector output (ADC counts, TDC, hit maps) into energy measurements, particle positions and directions and momentum.
 - Written in FORTRAN in previous runs.
 - Written in C++ in Run 2.
 - Hundreds of packages or modules, millions of lines of code, many 10's of authors.
 - An all volunteer army, with varying levels of software skills.



C++ Experience

- Big change from procedural to object-oriented language.
- Some resistance.
- Large training requirements.
- Need for C++ experts to support the physicists on design and coding.
 - Two individuals were hired by Fermilab to provide that support.
- The code runs, is probably as fast or faster than Fortran code, and in general the exercise has been successful.
- Most (not all) new experiments choose C++ for offline event reconstruction.



Was the transition to C++ beneficial?

- I'm not an expert and haven't worked with the code directly.
- The answer probably won't be known for some time:
 - Will code be more easily maintainable?
 - Will the code be more robust?
 - Will the code be as fast or at least not too slow?
 - Will we be aligned better with industry and other code developers?



Other software for Run 2

- Mixture of commercial, lab-developed and open source.
- Each product is chosen based on its ability to solve a problem and on its cost (both to write and to support).
- Long list of products, some examples:
 - Linux, gcc, emacs, MySQL
 - KAI C++ compiler, LSF (Batch system), Purify
 - FBS, Enstore, SAM, ftt, ZOOM
 - GEANT3/4, ROOT



The Future

- GRID Computing
- Killing Tapes.



Are Grids a solution?

Computational Grids

Les Robertson, CERN

- Change of orientation of Meta-computing activity
 - From inter-connected super-computers
... .. towards a more general concept of a
computational power Grid (The Grid – Ian Foster,
Carl Kesselman^{**})
- Has found resonance with the press, funding agencies

But what is a Grid?

*"Dependable, consistent, pervasive access to resources^{**}"*

So, in some way Grid technology makes it easy to use diverse, geographically distributed, locally managed and controlled computing facilities – as if they formed a **coherent local cluster**

^{**} Ian Foster and Carl Kesselman, editors, "The Grid: Blueprint for a New Computing Infrastructure," Morgan Kaufmann, 1999



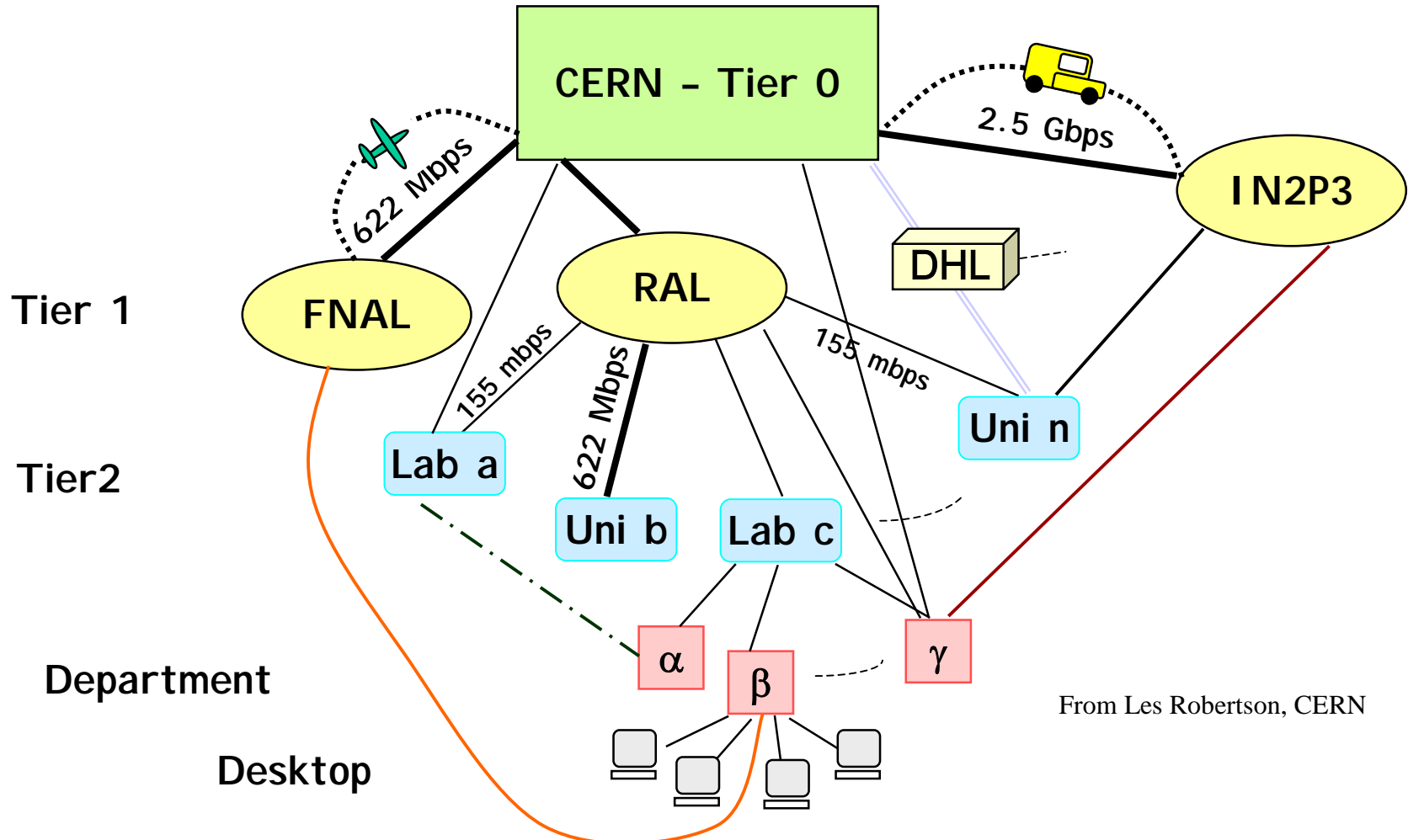
What does the Grid do for you?

Les Robertson

- You submit your work
- And the Grid
 - Finds convenient places for it to be run
 - Organises efficient access to your data
 - Caching, migration, replication
 - Deals with authentication to the different sites that you will be using
 - Interfaces to local site resource allocation mechanisms, policies
 - Runs your jobs
 - Monitors progress
 - Recovers from problems
 - Tells you when your work is complete
- If there is scope for parallelism, it can also decompose your work into convenient execution units based on the available resources, data distribution



CMS/ATLAS and GRID Computing



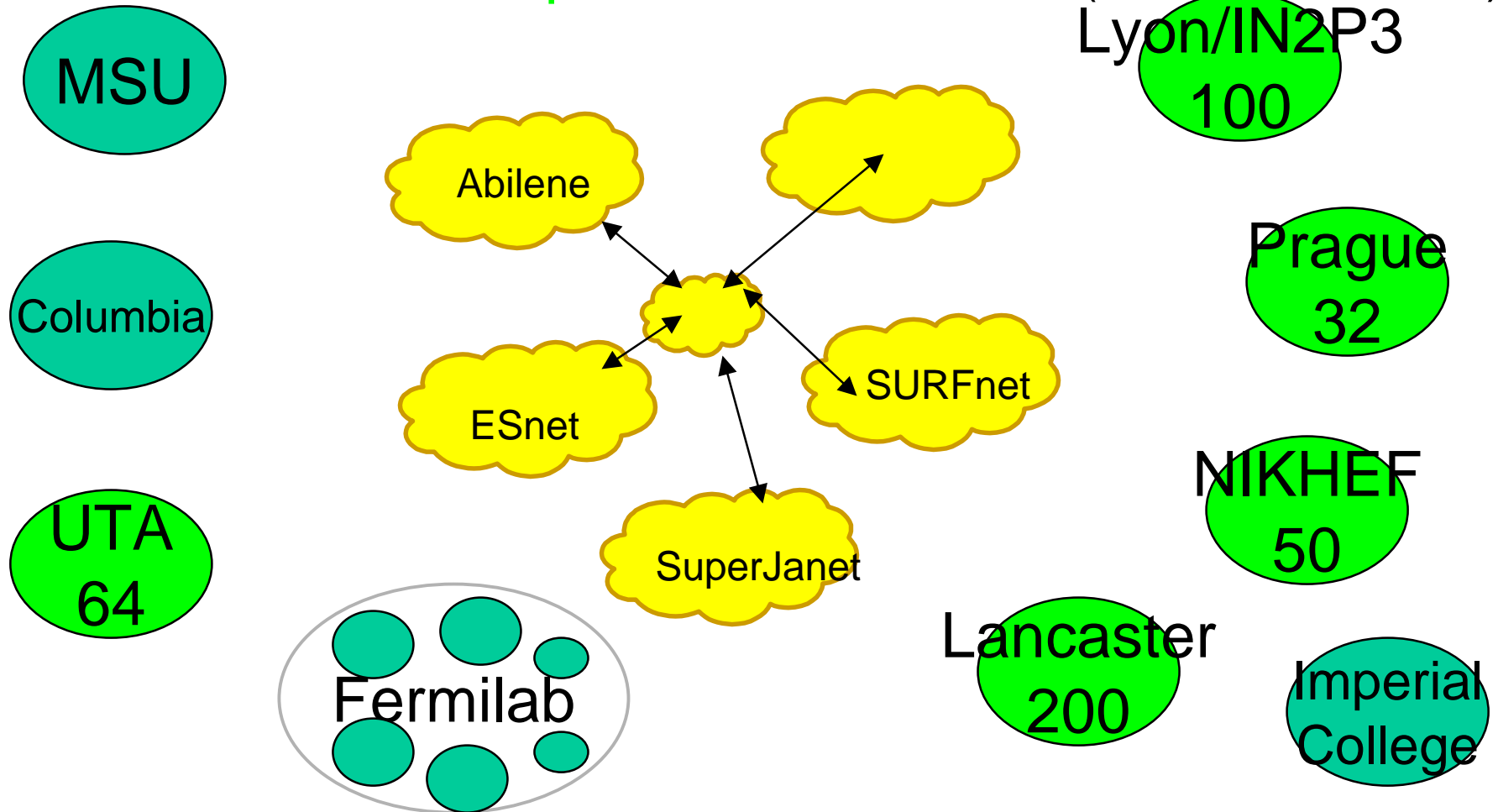
October 17, 2001

Stephen Wolbers ACM Chicago
October 2001



DO Processing Stations Worldwide

● = MC production centers (#nodes all duals)



October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

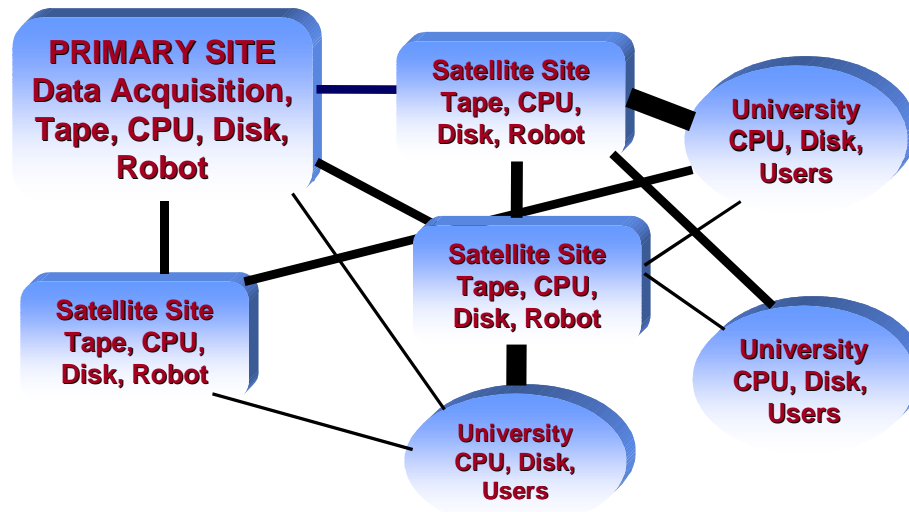


PPDG GRID R&D

Richard Mount, SLAC



PPDG Multi-site Cached File Access System



PPDG

November 15, 2000

LHC Computing Review

October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

50



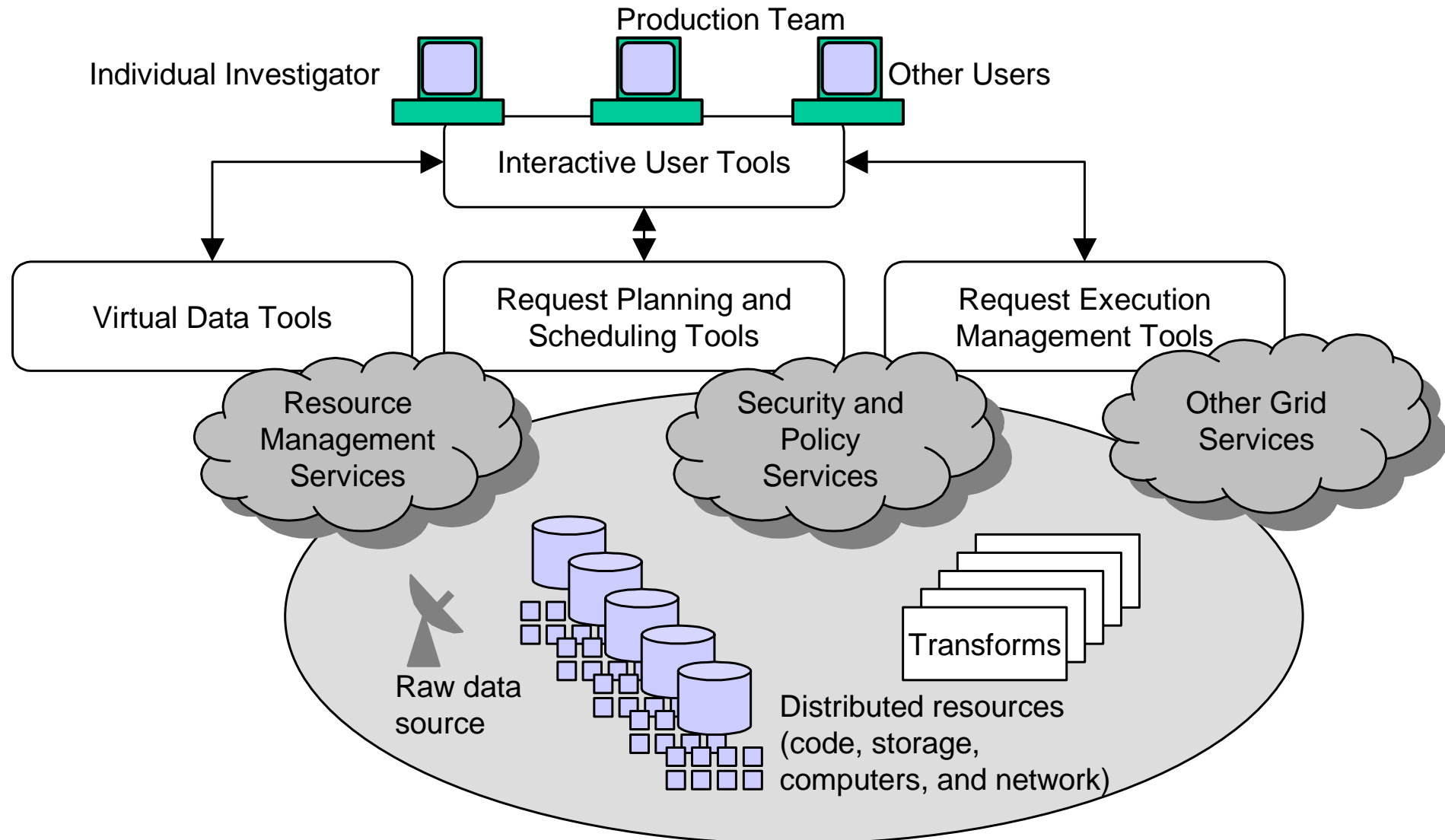
GriPhyN Overview

(www.griphyn.org)

- 5-year, \$12M NSF ITR proposal to realize the concept of virtual data, via:
 - 1) CS research on
 - Virtual data technologies (info models, management of virtual data software, etc.)
 - Request planning and scheduling (including policy representation and enforcement)
 - Task execution (including agent computing, fault management, etc.)
 - 2) Development of Virtual Data Toolkit (VDT)
 - 3) Applications: ATLAS, CMS, LIGO, SDSS
- PIs=Avery (Florida), Foster (Chicago)



User View of PVDG Architecture





GRID Computing

- GRID computing is a very hot topic at the moment.
- HEP is involved in many GRID R&D projects, with the next steps aimed at providing real tools and software to experiments.
- The problem is a large one and it is not yet clear that the concepts will turn into effective computing.



Killing Tapes

- Storing petabytes of data is not easy.
- Disk is too expensive (today) and difficult to manage at this scale.
- Tape is prone to failure and improvements are not keeping up.
- Nevertheless, we have to keep the data and tape is how we do it.



Robots and tapes



October 17, 2001

Stephen Wolbers ACM Chicago
October 2001

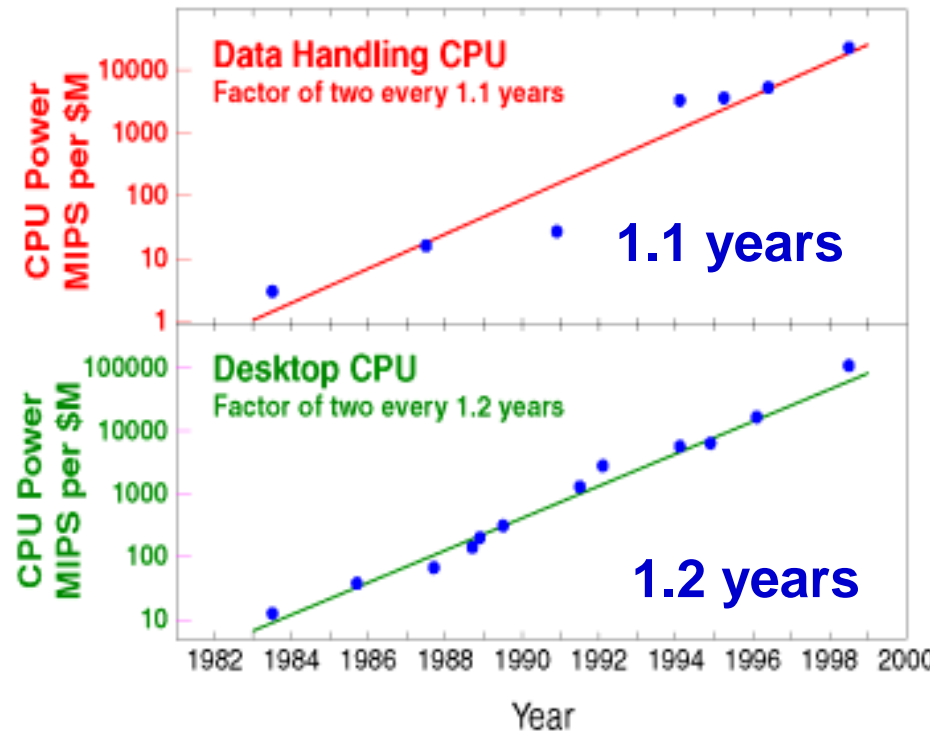
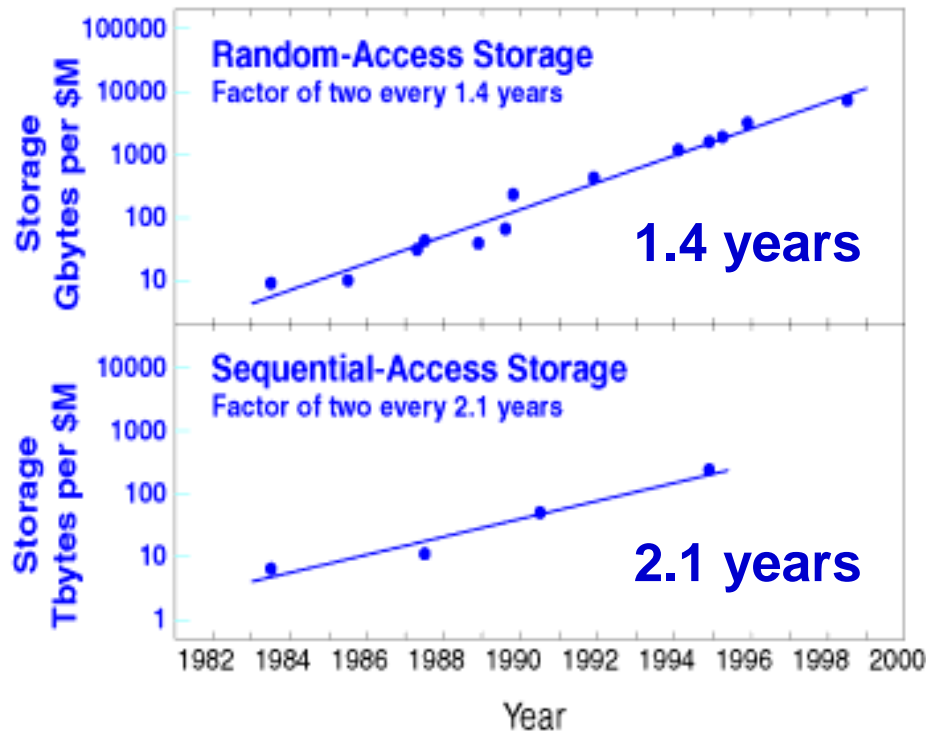
55

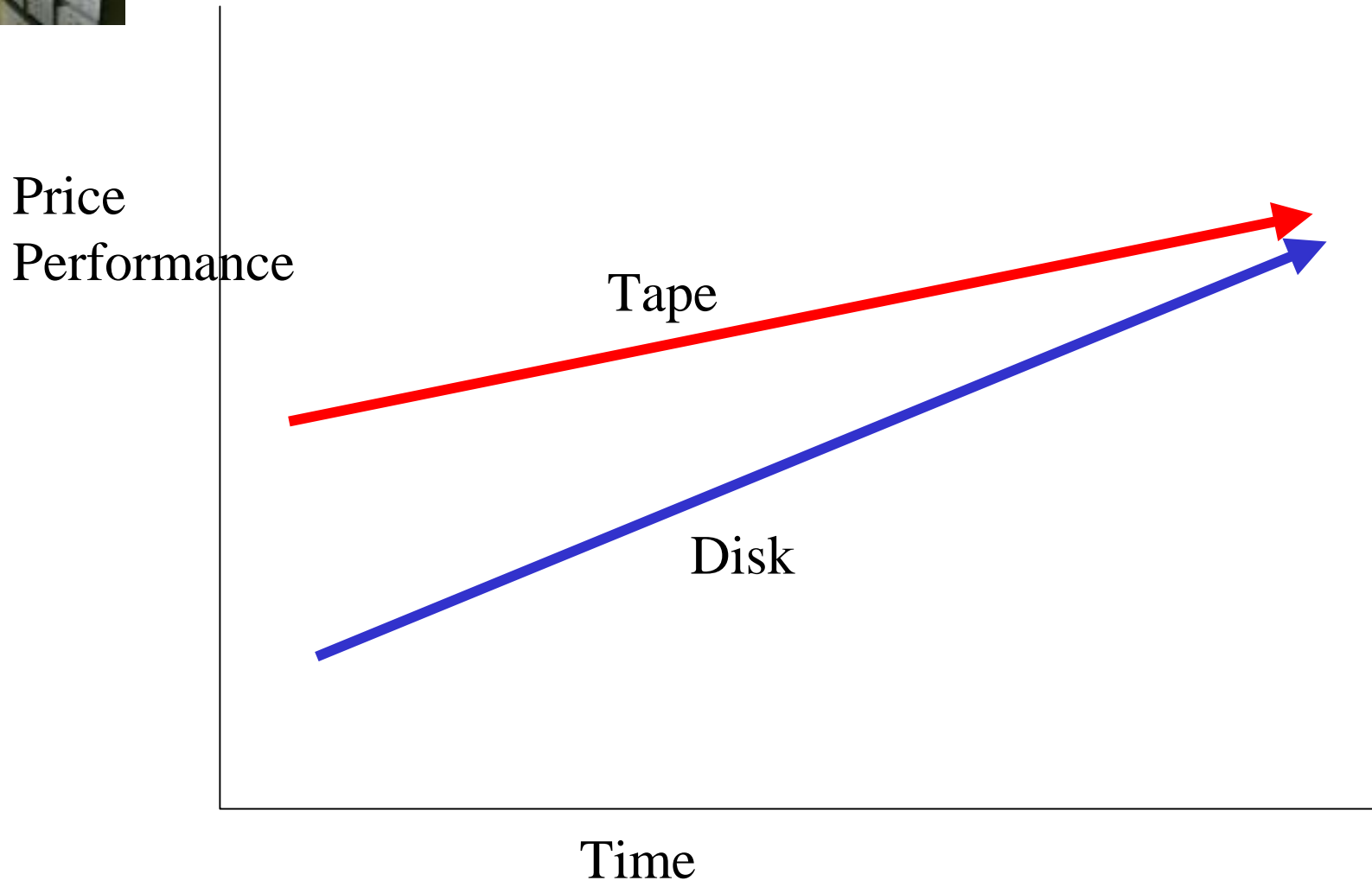


Hardware Cost Estimates

Paul Avery

Purchase Experience







An Idea: Disk Farms

- Can we eliminate tape completely for data storage?
- What makes this possible?
 - Disk drives are fast, cheap, and large.
 - Disk drives are getting faster, cheaper and larger.
 - Access to the data can be made via the standard network-based techniques
 - NFS, AFS, tcp/ip, fibrechannel
 - Cataloging of the data can be similar to tape cataloging



Disk Farms

- Two Ideas:
 - Utilize disk storage on cheap PCs
 - Build storage devices to replace tape storage
- Why Bother?
 - The price performance of disk is increasing very rapidly.
 - Tape performance is not improving as quickly.



I.-Utilize cheap disks on PCs

- All PCs come with substantial EIDE disk storage
 - Cheap
 - Fast
 - On CPU farms it is mostly unused
- Given the speed of modern ethernet switches, this disk storage can be quite useful
 - Good place to store intermediate results
 - Could be used to build a reasonable performance SAN



II.-Build a true disk-based mass storage system

- **Components of all-disk mass storage:**
 - Large number of disks.
 - Connected to many PCs.
 - Software catalog to keep track of files.
- **Issues**
 - Power, cooling.
 - Spin-down disks when not used?
 - Catalog and access of millions of files.



Summary

- Particle Physics and Computing are connected very deeply, and have been for decades.
- The biggest issue facing HEP is access to the huge datasets which are being generated.
- Clever ideas, new techniques and partnerships are welcome and needed to make progress.